Class Prediction and Discovery based on Gene Expression Data

Leping Li [1,*], Lee G. Pedersen[2,3], Thomas A. Darden[2] and Clarice R. Weinberg[1]

[1] Biostatistics Branch and [2] Laboratory of Structural Biology, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

[3] Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-3290

*Correspondence should be addressed to L.L. (voice: 919-541-5168, fax: 919-541-7880, email: Li3@niehs.nih.gov)

Keywords: k-nearest neighbors, genetic algorithm, gene selection, high-dimensional, and microarray

Abstract

Microarrays allow the expression patterns of tens of thousands of genes to be monitored in parallel. The technique has been used for gene expression profiling of normal and malignant cells. The major goal of these studies is to identify a subset of informative genes for class prediction as well as to uncover classes that were previously unknown (class discovery). The continued success of the methodology depends on the improvement of quantitative aspects of the microarray technology and on the development of computational tools that can mine the resulting large data sets.

Introduction

With large human expressed sequence tag (EST) sequence sets now available and with advances in microarray technology, it has become possible to monitor the genome-wide gene expression patterns of biological systems. Recently, microarray technology has been used to profile the global gene expression patterns of normal and transformed human cells in several tumors including colon (Alon *et al.*, 1999), leukemia (Golub *et al.*1999), prostate (Bubendorf *et al.*, 1999), breast (Perou *et al.*, 2000), lymphoma (Alizadeh *et al.*, 2000), and melanoma (Bittner *et al.*, 2000). Microarrays have also been used for gene expression profiling of the NCI's 60 tumor cell lines (Ross *et al.*, 2000). These studies may provide mechanistic insight about cell maltransformation and help in identifying biomarkers for cancer classification (molecular diagnosis).

While microarrays have been successfully used in gene expression profiling of tumor cells/tissues, successful application of the microarray technology in cancer classification may rely on data mining tools. This is because, of the many thousands of genes examined, only a fraction may present distinct profiles for different classes of samples (e.g. tumor vs. normal). Thus, it is critical to have computational tools that are capable of identifying a subset of informative genes embedded in a large data set that is contaminated with high-dimensional noise. Although many data analysis approaches have been proposed, few are designed specifically for class prediction and class discovery. Herein, we briefly review the structure of microarray expression data and these analysis methods.

Classification methods

Pattern recognition methods can be divided into two categories: supervised and unsupervised. A supervised method is a technique that one uses to develop a predictor or classification rule using a learning set with known classification. The predictor is subsequently used to classify unknown objects. Methods in this category include k-nearest neighbors (KNN) (Li *et al*., 2000a & 2000b), support vector machines (SVM) (Cortes & Vapnik, 1995), and linear discriminant analysis (LDA) (Vandeginste *et al*., 1998). Unsupervised pattern recognition largely refers to clustering analysis for which class information is not known or not required. Unsupervised methods include hierarchical clustering (Eisen *et al*., 1998), K-mean clustering (Tavazoie *et al*., 1999), and self-organizing map (Kohonen, 1999). Note that the KNN method can also be unsupervised.

Microarray data structure

1). Large number of genes and few samples

Unlike the conventional data sets that consist of a large number of observations (samples) and few parameters, microarray data consist of a large number of genes (parameters) and a small number of samples. When building a class predictor using a supervised pattern recognition method, many distinct class predictors may be obtained. In other words, several subsets of genes that can distinguish between different classes of samples may exist. Such subsets may be regarded as competing near-optimal solutions. It is possible that the solution space is relatively flat and that a "global" optimal solution may not exist. Consequently, it becomes important to identify many subsets of genes that can potentially discriminate between different classes of samples so that the relative importance of individual genes for sample classification can be assessed through statistical analysis of the near-optimal solutions.

2). Multivariate structures

In many classification problems, a multi-dimensional space is needed for sample separation. In a hypothetical

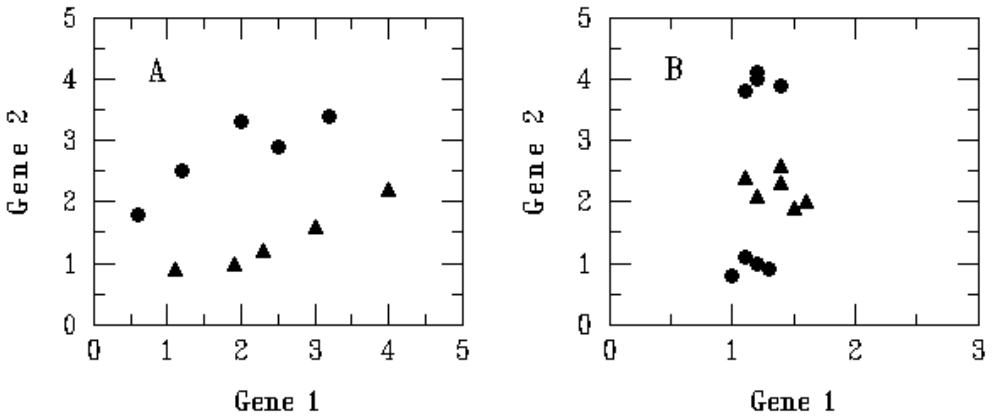

Figure 1. Scatter plots of hypothetical expression level of gene 1 vs. gene 2. Two different classes of samples are shown (filled cycles and triangles). Panel A illustrates that the two genes are *jointly* discriminative, but non-discriminative singly. Panel B shows that one class of samples (filled cycles) has two distinct subcategories. A simple approach like the *t*-statistic that treats these samples as a whole would fail to identify the two genes that are apparently discriminative.

example shown in Figure 1a, separation between the two classes of samples is apparent. However, considerable overlap occurs when the samples are projected onto either of the two dimensions. Thus, the two genes are *jointly* discriminative, but not discriminative singly. This illustrates that a set of genes should be considered simultaneously for their *joint* ability to discriminate, not individually. Recently, multivariate approaches such as principal component analysis (PCA) have been used in tumor classification (Alaiya *et al.*, 2000) based on gene expression data. Also, Kim *et al.* (2000) has used a multivariate approach to elucidate gene relationships.

3). Sample heterogeneity

Intuitively, a simple approach such as the student *t*-statistic may be applied to identify the differentially expressed genes. Essentially, the *t*-statistic approach searches for genes that deliver the largest difference in average intensity (expression level) between different classes of samples (e.g. normal vs. tumor samples) and the smallest variation within each class. While it is reasonable that genes with a large *t*-statistic would be discriminative, genes with a small magnitude *t*-statistic may also be discriminative (Fig. 1b). For instance, certain genes could be highly differentially expressed in one subcategory of a tumor but not in another. Samples of the same class may be at different developmental stages. Such genes that could well be informative may not be identified using the *t*-statistic as the selection criterion, since the *t*-statistic could be small. Scenarios of this nature are not unlikely, since tumors can be heterogeneous (Lengauer *et al.*, 1998). Thus, computational tools that can work well in the presence of sample heterogeneity are preferred. Subsequently, hidden subcategories in the sample may be discovered and may prove to be etiologically or prognostically important.

Class prediction

Methods for selecting a subset of informative (discriminative) genes for sample classification have recently been proposed (Golub *et al.*, 1999, Ben-Dor *et al.*, 2000, Li *et al.*, 2000a & 2000b). Golub *et al.* (1999) successfully applied neighborhood analysis to identify a subset of genes that discriminates between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), using a separation measure similar to the *t*-statistic. The 50 genes that best distinguish AML from ALL in 38 training set samples were chosen as a class predictor that correctly classified 36 of the 38 training set samples. When these genes were subsequently used to predict the class membership of new leukemia cases, 29 of the 34 test set samples were correctly classified with high confidence.

Recently, Ben-Dor *et al.* (2000) applied several classification methods (both supervised and unsupervised) to a colon (Alon *et al.*, 1999) and an ovarian data set including the KNN (without gene selection) and SVM (after gene selection). A boosting technique (Freund & Schapire, 1997) was used to search for a threshold (expression level) for each gene that would maximally discriminate between two types of samples (e.g. normal vs. tumor). Those that gave the smallest classification errors were taken as the relevant genes. All samples were used to obtain the relevant genes using the *leave-one-out cross-validation*

procedure as a measure of prediction strength (no test set was employed).

Although differing in technical details, both approaches (Golub *et al.*, 1999, Ben-Dor *et al.*, 2000) identify informative genes by examining one gene at a time (univariate), that is, samples were compared in single dimensions. Furthermore, both approaches implicitly assume that genes are similarly expressed within each type of sample. This could be problematic if subtypes exist so that the relevant genes are not uniformly expressed (as in Figure 1b).

Recently, we proposed a multi-dimensional classification method that not only selects a small subset of genes that *jointly* discriminate between different classes of samples, but also assesses the relative predictive importance of all genes for sample classification. Details of the method (Li *et al.*, 2000a) and a study of the sensitivity to choices of various parameters of the method have been reported (Li *et al.*, 2000b). In brief, the method employs a non-parametric pattern recognition approach, the k-nearest neighbors (KNN), and a searching tool, a genetic algorithm (GA). The GA is used to search high-dimensional space, since selecting a subset of genes from a large gene pool is a combinatorial problem. For instance, the number of ways of selecting 50 genes from 2000 is approximately 10^{100}. The KNN method is used as the classification tool that distinguishes between discriminative and non-discriminative genes. Simply speaking, we employ the GA to choose a relatively few subsets of genes (from many combinations) for testing with KNN as the evaluation tool. The GA/KNN method searches for many subsets of genes that potentially discriminate between different classes of samples using the training set. When many such subsets of genes have been obtained, the relative importance of genes for sample classification is assessed by examining the frequency of gene memberships in those near-optimal subsets. The genes can then be ranked based on the frequency of their selection. We divide the data sets, whenever allowed, into a training and a test set. The training set is used to build a class predictor while the test set is used to validate the class predictor.

We have applied the GA/KNN method to colon cancer data (Alon *et al.,* 1999), lymphoma data (Alizahel *et al.*, 2000) (http://llmpp.nih.gov/lymphoma/), and leukemia data (Golub *et al.,* 1999) (http://llmpp.nih.gov/lymphoma/). The results have been reported (Li *et al.*, 2000a & 2000b) and are available on the web (http://chun.nihes.nih.gov/~leping/). For all data analyzed, "unknown" samples (in the test set) were largely classified correctly using the 50 top-ranked genes. For the colon data set, the GA/KNN method found that three of the tumor specimens (T30, T33, and T36) and two of the normal specimens (N34 and N36) were predicted to be in the wrong set. This prediction was, however, later confirmed using pathology and the anomalies appear to have resulted from sample contamination (personal communication with Dr. Uri Alon). For the leukemia set (Golub *et al.*, 1999), the GA/KNN correctly classified all training set samples and all but one of the 34 test set samples (AML-66) using the 50 top-ranked genes. Furthermore, the set of discriminative genes revealed the existence of two subtypes within the ALL class without applying prior knowledge.

Class discovery

Current computational tools for class discovery based on gene expression data have been largely limited to clustering analysis. In a paper by Alizadeh *et al.* (2000), cDNA microarrays were used for gene expression profiling of a set of normal and malignant lymphocyte samples. Hierarchical clustering analysis (Eisen *et al.*, 1998) suggested that B-cell differentiation genes may be used to subdivide diffuse large B-cell lymphomas (DLBCL). Subsequent clustering analysis using those genes has led to the discovery of two distinct DLBCLs: germinal center B-like and activated B-like DLBCL (Alizadeh *et al.,* 2000). The subclassification of DLBCL appears to correlate with the overall survival of the patients (Alizadeh *et al.,* 2000). Similar profiling studies on cutaneous malignant melanoma (Bittner *et al*., 2000) and breast tumors (Perou *et al*., 2000) have been reported.

In addition to AML and ALL class prediction, Golub *et al.* (1999) also applied self-organizing maps (SOM) for class discovery. Distinct subcategories of AML and ALL were identified. Similarly, the GA/KNN method (Li *et al.*, 2000a) was also able to uncover clinically distinct subtypes within ALL without prior knowledge.

Conclusions

Gene profiling has been shown to be promising in aiding cancer classification. Methods for class prediction and class discovery based on gene expression data have been proposed (Golub *et al.*, 1999, Ben-Dor *et al.*, 2000, Bittner *et al.*, 2000, Li *et al.*, 2000a & 2000b, Perou *et al.*, 2000). Most of the approaches utilize methodologies that were developed many years ago. These methods have been shown to be useful in mining complex microarray data. Regardless of the pattern recognition method used, it is important to understand the strengths and limitations of the method as well as the data structure to which the method is applied. It may not be difficult to construct a class

predictor when sample class membership is known. Supervised pattern recognition methods may be applied.

It is much more challenging if the class membership is unknown while simultaneously determining a class predictor (class discovery). Currently, the main approach used for class discovery is clustering analysis using either a subset of genes or all genes. Generally, a subset of informative genes, instead of all genes, should be used for class discovery, since not all genes are relevant to class distinction. Without knowing the classifications, it is difficult to identify the subset of informative genes. A solution to such a problem is possible and work is in progress. The methods ultimately developed should be useful not only for cancer data sets, but also for those from environmental or pharmaceutical studies for which class membership is not totally clear. In the latter case, compounds of the same class may not behave in the same manner. The class relationship may well change as a function of dose and/or exposure time. Thus, methods that can identify a small subset of (signature) genes that can *jointly* distinct one class from another should be valuable.

As the microarray technology becomes more quantitative and the computational tools that mine the resulting large data sets become mature, tumor classification and class discovery using gene expression profiling may revolutionize cancer biology.

References

Alaiya, A.A., Franzen, B., Hagman, A., Silfversward, C., Moberger, B., Linder, S. and Auer, G. *Int. J. Cancer*, **2000**, *86*, 731.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr, Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, E., Grever, M.R., Byrd, J.C., Botstein, D., Brown and P.O., Staudt, L.M. *Nature*, **2000**, *403*, 503.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. *Proc. Natl. Acad. Sci. USA,* **1999***, 96*, 6745.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. In *Proceedings of the Fourth International Conference on Computational Molecular Biology (RECOMB2000)*, ACM press, New York, **2000**.

Bittner, M., Meitzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N. and Trent, J. *Nature*, **2000**, *406*, 536.

Bubendorf, L., Kolmer, M., Kononen, J., Koivisto, P., Mousses, S., Chen, Y.D., Mahlamaki, E., Schraml, P., Moch, H., Willi, N., Elkahloun, A.G., Pretlow, T.G., Gasser, T.C., Mihatsch, M.J., Sauter, G., Kallioniemi, O.P. *J. Nat. Cancer Inst*., **1999**, *91*, 1758.

Cortes, C. and Vapnik, V. *Machine Learning*, **1995**, *20*, 273.

Freund, Y., Schapire, R.E. *J. Comput. Sys. Sci*., **1997**, *55*, 119.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H, Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. *Science*, **1999**, *286*, 531.

Kim, S.C., Dougherty, E.R., Chen, YD., Sivakumar, K., Meltzer, P., Trent, J.M. and Bittner, M. *Genomics*, **2000**, *67*, 201.

Kohonen, T. *Self-Organizing Maps*, Springer, Berlin, **1995**.

Lengauer, C., Kinzler, K.W. and Vogelstein, B. *Nature*, **1998**, *396*, 643.

Li, L., Darden, T.A., Weinberg, C.R. and Pedersen, L.G. *Combinatorial Chemistry & High Throughput Screening*, **2000a**, accepted.

Li, L., Weinberg, C.R., Darden, T.A. and Pedersen, L.G. *Bioinformatics*, **2000b**, submitted.

Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Aksien, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Børresen-Dale, A.L., Brown, P.O. and Botstein, D. *Nature*, **2000**, *406*, 747.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C.E., Lashkari, D., Shalon, D., Myers. T.G., Weinstein, J.N., Botstein, D. and Brown, P.O. *Nature Genet*., **2000**, *24*, 227.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. *Nature Genet*., **1999**, *22*, 281.

Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J. In *Handbook of Chemometrics and Qualimetrics, Part B,* Elsevier Science, The Netherlands, **1998**.